U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Ocean Service
Office of Response and Restoration
Coastal Protection and Restoration Division
c/o EPA Region X (ECL-117)
1200 Sixth Avenue
Seattle, Washington 98101

June 30, 2006

Eric Blischke
U.S. Environmental Protection Agency
Oregon Operations Office
811 SW Sixth Avenue
Portland, OR 97204

Chip Humphrey
U.S. Environmental Protection Agency
Oregon Operations Office
811 SW Sixth Avenue
Portland, OR 97204

Dear Chip and Eric:

This letter provides **NOAA's comments on the Portland Harbor Superfund Site Ecological Risk Assessment *Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests (Draft)***. The document, prepared by Windward Environmental LLC for the Lower Willamette Group (LWG), is dated March 17, 2006.

NOAA appreciates this opportunity to provide comments on this draft report and would like to commend the LWG on the tremendous amount of effort that went into its preparation. In general, NOAA believes that LWG's proposed approach will serve as a useful tool in assessing risk and informing remedial decision making at the Portland Harbor site. However, NOAA does have some concerns and these are outlined in more detail in the sections that follow. It is NOAA's recommendation that these concerns should be carried forward and addressed, as discussed below, in the Round 2 Comprehensive Approach Report.

In short, our primary concerns are as follows: 1) The proposed threshold numbers for PAHs as derived from the Floating Percentile Model (FPM) are patently unacceptable and should be discarded. Such values for PAHs should be based on the Logistic Regression Model (LWM) and, as appropriate, existing literature and/or relevant regulatory/screening level values. 2) As implied above, the LRM should *not* be discarded as an interpretive tool. A preliminary analysis of the results of both models suggests significant overlap in results (i.e., in the delineation of areas of no-risk vs. risk) which may help to focus additional lines of evidence on areas where modeled risk to benthos is uncertain. Results of both the FPM and LRM should be carried forward. 3) The omission of the *Hyalella* growth endpoint from the FPM is not appropriate (see comment below). This endpoint should be carried forward and

FPM results including this endpoint should be presented in the Round 2 Comprehensive Approach Report.

General Comments

***Hyalella* growth and survival endpoint**: NOAA notes that the LWG proposes to disregard the results of the *Hyalella* growth and survival (pooled) endpoint. LWG supports this proposal based on "difference from other endpoints" and "no correlation with mortality endpoint". Yet these are precisely the reason that multiple test endpoints are required (because different test endpoints may show different sensitivities to different chemical mixtures). However, there was substantial agreement between the *Hyalella* and *Chironomus* pooled endpoints for samples that showed an extreme degree of toxicity (eg., < 50% of control) in either test. The "lack of correlation to Chemicals of Concern" and the "effect of percent fines" may be more related to the different contaminant mixtures and gradients in the Portland Harbor study area. In a complex environment with multiple chemical mixtures and gradients with limited numbers of samples from any one area, a lack of correlation between a test endpoint and individual chemicals does not necessarily imply that toxicity is not related to chemical contamination. This is supported by the differences in chemicals that "set" the different models for the same sample (for example, the chemical with highest ratio of concentration to floating point value for a sample may be a phthalate, while the chemical with the highest probability of toxicity in logistic regression models may be ammonia or DDT for the *Hyalella* pooled model or PCBs or cadmium for the *Chironomus* pooled model). Because each contaminant can be considered as an indicator of toxicity for the chemical mixtures, it is not surprising that generic indicators such as percent fines, ammonia, or sulfides are good predictors of toxicity.

Proposed total PAH threshold values: The proposed Effects Level 2 and Effects Level 3 concentrations for total PAH, which represent AET values, are unreasonably high (1270 ppm DW) and significantly higher than other published values. For example, the proposed value exceeds the consensus-based freshwater PEC for Total PAH (22.8 ppm DW; MacDonald et al 2000) by more than a factor of 50. Of the samples exceeding the PEC value, 73% have a Level 2 response or greater in one or both of the pooled endpoints and 86% for samples with at least 25% fines. If we exclude the *Hyalella* growth endpoint, 62% of the samples exceeding the PEC have Level 2 or greater response compared to 65% of the samples with diesel concentrations exceeding the proposed FPM value of 340 ppm. While diesel concentrations may be a slightly better predictor of toxicity than total PAH for this dataset, total PAH concentrations much lower than the proposed AET values are reliable predictors of toxicity. NOAA considers that the proposed values for total PAH serve no useful purpose and should be discarded.

Level 1 Biological Effects Level: The report states "*it is recommended that Level 1 not be used to set SQVs for Portland Harbor because it is relatively unreliable in accurately predicting effects and well below the cleanup levels set at other regional Superfund sites.*" NOAA agrees that Level 1 Biological Effects Level values should not be used as target cleanup levels. However, Level 1 values should not be discarded, as they represent

concentrations associated with low level effects and provide useful information for defining areas of concern. The incidence of Level 1 or greater effects increases with increasing probability of toxicity.

Single-threshold evaluation of reliability: The report relies exclusively on a single-threshold evaluation of "reliability" of sediment quality guidelines. The conceptual model that a single value can accurately distinguish between "good" and "bad" samples, while perhaps desirable, is not consistent with most environmental data. NOAA agrees that minimizing false negatives and false positives is an important goal, but concentration-response relationships are usually continuous and multiple thresholds may provide better separation of false positive and negative concentrations. For continuous models, such as the logistic regression model, an evaluation based on a single-threshold loses important information.

LRM model development: The logistic regression models were developed following the published approach developed by NOAA and EPA (Field et al. 1999; Field et al 2002; EPA 2005). The model development presented in the report did not address exclusion of chemical models that resulted in a high degree of false positives or adjustments to the screening approach to reduce the influence of a small number of non-toxic samples with very high chemical concentrations, which was particularly problematic for PAHs. The models were evaluated for reliability using the single threshold approach. NOAA recognizes that this evaluation provides some useful information, but reducing the evaluation to a single threshold does not take full advantage of the continuous concentration-response relationship.

NOAA developed alternative logistic regression models, using a larger freshwater database for the *Hyalella* 28-day growth and survival endpoint and calibrated these models to the Level 2 Effect Level in the Portland Harbor data.

Recommended FPM values: The recommended FPM values are based on 3 individual endpoints (*Chironomus* survival, *Chironomus* growth, and *Hyalella* survival), excluding results for the *Hyalella* growth endpoint and for the combined (pooled) growth and survival endpoints for both test species. The pooled results are important to consider, because growth and survival are not independent measures. (See previous discussion of the rationale for including the *Hyalella* growth and survival combined endpoint.)

Several of the recommended FPM values have the same concentration for Level 2 and Level 3 Effects. This indicates that these values are at the upper end of the concentration-response relationship and thus may be considered extreme effect concentrations.

PEC-quotient approach: The report did not evaluate the PEC-quotient (PEC-q) approach (Ingersoll et al 2001) – one of the major approaches to developing freshwater guidelines – which has been applied effectively in other Superfund remedial investigations (e.g., Calcasieu Estuary, Louisiana). A quick review of the data indicate that samples with mean PEC-q's greater than 1 show a Level 1 response or greater in at least one toxicity test endpoint in 87% of the samples and at least a Level 2 response in 77% of the samples. This

suggests that the PEC-q approach may be useful in contributing to the identification of areas of concern. NOAA recommends that evaluation of the Ingersol PEC-q would help determine its potential useful for the Portland Harbor remedial investigation.

2.0 DATA QUALITY AND ORGANIZATION

Page 5 "*petroleum data for 203 stations*" How were the 146 stations with matching toxicity data for petroleum analysis selected?

Page 5 The report states that "*The biological effects levels used in the analyses are intended to correspond conceptually to "no effects level" (Level 1), "minor effects level" (Level 2), and "moderate effects level" (Level 3). As requested by EPA (EPA 2005a), the three levels were set at 90, 80, and 70% of the response observed in the control sediment, respectively.*" The biological effect levels are mischaracterized. A more appropriate characterization would be "minor effects level" (Level 1), "moderate effects level" (Level 2), and "severe effects level" (Level 3). NOAA recommends that the effects level characterizations be revised accordingly.

Page 8 The report states that "*The exclusion of data with the N-qualifier primarily affected the pesticide data. Between 23 and 53% of the data for the following pesticides were excluded: aldrin, hexachlorocyclohexane (alpha-, beta-, and delta-), nonachlor (cis- and trans-), dieldrin, and methoxychlor. Between 35 and 67% of the summed data of DDD, DDE, DDT, total DDT, total chlordane, and total endosulfan were excluded.*" Considering that some of these contaminants are known to be of importance in the Lower Willamette, NOAA is not entirely clear on the implications for the aforementioned analysis. What percentage of the excluded data had concentrations that exceeded the 25th percentile of the detected/included data? Would including these data affect the results? NOAA requests clarification.

Page 9 The report states that "*The presence of non-toxic, naturally occurring crustal elements such as aluminum and selenium can confound the development of meaningful SQVs for the remainder of the analytes.*" It is not clear why this should be the case. This may be an issue for FPM development, but LRMs are developed independently for each chemical and the crustal elements can be included or not in the development of the maximum probability model. NOAA requests clarification and additional explanation.

Page 11 The report states that "*Individual dioxins and furans (replaced by TEQ).*" TEQs are based on tissue concentrations and are not meaningful in sediment without accounting for differences in bioaccumulation factors for individual PCB, dioxin, and furan congeners.

Page 11-12 The report states that "*Using summations reduces covariance problems, and past side-by-side comparisons of other Oregon and Washington data sets have shown better*

reliability when summations are used." Please provide reference(s) in support of this statement.

Page 12 The report states that "*Normalization of non-polar organic compounds and metals could be applied in an attempt to improve the reliability of the predictive model(s). However, no actual advantage has been revealed in past side-by-side comparisons of other Oregon and Washington data sets, and the reliability of the non-normalized sediment quality guidelines is generally the same or better than the normalized guidelines*." Please provide reference(s) in support of these statements.

3.0 COMPARISON TO EXISTING SEDIMENT QUALITY VALUES

Evaluation of the performance of paired values, such as TELs and PELs, using a single threshold is inappropriate. These types of sediment quality guidelines were developed to provide a lower level below which toxicity would be unlikely and a higher level above which toxicity would be likely.

"*Quotient Methods – Quotient methods were developed as an approach to increase the predictive ability of certain SQVs (Long et al. 1998)*" Please refer to and cite the key papers on development and application of freshwater quotients (Ingersoll et al 2001; MacDonald et al 2000). NOAA suggests that it would be useful to apply the PEC-q method presented by Ingersoll and MacDonald to the Portland Harbor data.

Page 16 The report states that "*In general, the quotient methods are an improvement over most of the SQV sets discussed above although not sufficiently reliable for use in predicting toxicity results at this site (see Appendix A). It is possible that the quotient approach has merit, but it needs to be optimized on a site-specific basis.*" A quick review of the data indicate that samples with mean PEC-q's greater than 1 show a Level 1 response or greater in at least one toxicity test endpoint in 87% of the samples and at least a Level 2 response in 77% of the samples. This suggests that the PEC-q approach may be useful in the identification of areas of concern. NOAA requests that the LWG present the results of the PEC-q analysis conducted by LWG for this report.

4.0 EXPLORATORY ANALYSES TO SUPPORT DEVELOPMENT OF SITE-SPECIFIC SQVS

Page 17 Please explain what is meant by the term "*chemical endpoints*"?

Fig. 4-1 is not clearly explained. For example, it is unclear whether everything was correlated with everything in the table and only the highest correlations identified.

Page 18 The report states that "*Even if correlations were not highly linear throughout the range, it was true for nearly all chemicals that high concentrations occurred in sediments with the highest fine-grained fractions (i.e., high concentrations implied high percent fines, but high percent fines did not always imply high concentrations).*" This also implies that, in

general, high percent fines are a good indicator of high chemistry and that low percent fines are good indicator of low chemistry.

5.0 DEVELOPMENT OF BENTHIC TOXICITY PREDICTION MODEL

Page 23 The report states that "*These ranges may overlap due to site-specific or sample-specific variations in bioavailability or toxicity*." This statement appears to assume causality, which may not be the case. The concentrations for a chemical that are associated with toxicity may have at least as much to do with the mixtures of other chemicals present in the sample as bioavailability.

"*...and this is the source of most of the false positive errors*." NOAA is not clear on what is meant by this statement or where it is shown. Please provide clarification.

"*Above the red bar, both false negatives and false positives may occur, as is shown for Chemicals A, B, and C. This region is the range of concentrations over which sample-specific bioavailability plays an important role in toxicity,...*" Please explain the basis for the bioavailability assertion. Does this assume causality for individual chemical concentrations?

Page 24 The report states that "*...hand-optimization steps were used to identify chemical concentrations for each endpoint and effects level in order to minimize prediction errors*." Please explain further how this was accomplished?

Page 26 The report states that "*Certain chemicals had no significant differences for any of the hit/no-hit definitions or endpoints. These included: 4-methylphenol, aldrin, alpha-hexachlorocyclohexane, antimony, bis(2-ethylhexyl)phthalate, butylbenzyl phthalate, chromium, delta-hexachlorocyclohexane, dibutyltin, hexachlorobenzene, monobutyltin, pentachlorophenol, phenol, tetrabutyltin, total dioxins/furans, total endosulfans, and tributyltin*." It appears that this statement is not consistent with the results in Table 5-2 for at least 4-methylphenol, antimony, and pentachlorophenol. Please check and revise accordingly or provide clarification.

Page 29 The report states that "*It is also interesting to note that for most endpoints, bulk petroleum (diesel-range hydrocarbons and residual-range hydrocarbons) was somewhat more strongly correlated with toxicity than were total PAHs, in spite of the fact that PAHs were measured at all stations, and bulk petroleum was measured at only a subset of stations*." Diesel- and residual-range hydrocarbons were only measured at selected stations. What was the basis for selecting the stations for the petroleum hydrocarbon analysis? For the stations selected for hydrocarbon analysis, diesel and total PAH were strongly correlated. [the average total PAH concentration was much higher for samples with diesel measured, approximately 126 ppm compared to 2.6 ppm for the other samples.]

Page 36 The report states that "*...there are a limited number of analytes for which FPM values can be calculated because the level at which these analytes reach their toxicity threshold is apparently above their concentration ranges in this data set*." The term

"toxicity threshold" appears to assume causality for an individual chemical. In environmental mixtures, this is an unjustified assumption.

Page 53 The report states that selection of a single threshold from a continuous relationship is not a useful application of these models.

Page 55 Regarding the reference to "*Chemical drivers*", please clarify that "chemical drivers" refers only to chemicals that play a role in the predictive model (i.e., the best predictors of toxicity of the chemical mixtures in the study area) and may have nothing to do with "chemical drivers" of toxicity.

Page 56 The report states that "*An effect of grain size on toxicity is seen only for Hyalella pooled at Levels 2 and 3. This correlation between the Hyalella pooled and percent fines is indicated by the presence of percent fines as a chemical driver*." A correlation with percent fines does not demonstrate a grain size effect and does not imply that percent fines is causing toxicity. The highest concentrations for each chemical are associated with samples with high percent fines, so it cannot be concluded that fines are causing toxicity in the *Hyalella* pooled endpoint. (See next comment about the use of the term "chemical drivers".)

Page 56 Regarding the reference to "Chemical Drivers": Chemicals that are good predictors in the models should not be assumed to be causing toxicity. The report should make a clear distinction between chemicals that are "drivers" in the models and those that are associated with causality. Please revise accordingly.

Page 57 The report states that "*Ammonia and sulfides are common confounding factors in bioassays (ASTM 2003) and can sometimes be high enough to cause toxicity in bulk sediments, even when their levels in overlying water are below bioassay QA/QC criteria*." Please clarify the basis for the statement in the 2nd part of this sentence. Does information exist which shows that the bioassay QA/QC criteria values for ammonia and sulfides in overlying water are too high?

6.0 SUMMARY AND CONCLUSIONS

Page 58-59 The report states that "...*it became clear that the Hyalella growth endpoint was responding differently than the other endpoints from a variety of standpoints, which raised some concerns*." Isn't this a primary reason for using different toxicity endpoints?

Page 59 Regarding the reference to "*Effect of Percent Fines*": A correlation does not demonstrate an effect. As pointed out earlier, most of the high chemistry was found in high percent fines samples. Please change "effect of" to "correlation with" or similar term that does not imply causality.

"*Certainly, there are precedents for high- and low-percent fines effects on other amphipods, both freshwater and marine, in commonly used toxicity tests*." Please provide reference sources for this statement.

Page 60 **Level 1 Biological Effects Level**: "The reliability of nearly all the endpoints at Level 1 is reduced as compared to Levels 2 and 3. This is likely due to the very small difference (10%) from control used to define the Level 1 endpoints. This level of difference is likely within natural and laboratory variability in many cases"...A difference of 10-20% from control was statistically determinate for most of the samples for all endpoints, indicating that it was outside the range for laboratory variability for the tests conducted. The Level 1 Biological Effects Level is useful for identifying concentrations at the lower end of the concentration-response relationship, in contrast to the Level 3 concentrations, which are at the upper end of this relationship.

Page 62-63 "***Sensitivity to individual chemicals varies by endpoint****. The chemicals that showed a relationship to toxicity varied by endpoint. The Chironomus growth, Chironomus mortality, and Hyalella mortality endpoints were sensitive to similar chemicals, while the Hyalella growth endpoint showed a very different relationship*." Individual chemical sensitivity should not be asserted or implied from correlations with environmental chemical mixtures. NOAA suggests using terminology that refers to the relationship between toxicity endpoints and chemical concentrations as "correlation" or "association."

Page 63 The report states that "*The results of this model correspond well both with measured toxicity and with the conceptual site model*." NOAA is not clear on the meaning of this statement. In what way or how does the model correspond well with measured toxicity and the conceptual site model? Does this mean the model corresponds well with measured toxicity and those locations where one would expect to see toxicity based on the conceptual site model? Which conceptual site model(s) (ecological, human health, overarching CSM)? Please clarify.

Page 65 The report states that "*Bulk petroleum measures were more strongly correlated with toxicity than total PAHs, even though PAHs were measured at all stations, and bulk petroleum was measured at only a subset of stations. Although the SQVs for PAHs may appear high, they are consistent with those derived from other West Coast data sets (e.g., San Francisco Harbor (Germano & Associates 2004), Los Angeles Harbor (unpublished)) using the FPM and the LRM, indicating that PAHs alone are not large contributors of toxicity to benthic organisms. PAHs are only a small subset of the suite of narcotic chemicals present in sediments and in petroleum, all of which may affect benthic organisms through similar toxicological pathways (McCarty 1991; McCarty and Mackay 1993; McCarty et al. 1992). The bulk measures of petroleum appear to better capture and correlate with that toxicity, as is apparent from the SQVs calculated for these measures.*" In Los Angeles Harbor as well as the entire California Sediment Quality Objectives database, the total PAH concentrations were much lower – very few samples exceeded the ERM of 44 ppm and none were within an order of magnitude of the proposed values. The LRM results for Los Angeles Harbor showed that PAHs infrequently had the maximum probability for a sample, but the logistic regression model probability of toxicity associated with the proposed PAH SQV would be very close to 1 (maximum possible). NOAA is concerned that the

statement, as presented, is inaccurate and/or incorrect. NOAA is adamant that the presented SQVs for PAHs are not acceptable.

The report states that "*The FPM often identifies similar values for different effects levels, as can be seen in Table 6-1 (this is also true of AETs). Some chemicals, such as ammonia, arsenic, and residual-range hydrocarbons, have different SQVs at Level 2 and Level 3. Other chemicals, such as copper, diesel-range hydrocarbons, and DDTs, have the same SQV at both levels. Although at first this may appear unusual, it reflects the fact that the concentration-toxicity curve for these chemicals is apparently steep in Portland Harbor.*" Please provide the factual basis for this statement? Consider that this result may be interpreted to suggest that the similar values for different effects are near the upper end of the concentration-response relationship. This is certainly the case for total PAH.

Appendix A

A.4: "*For each existing SQV set, the more protective of the two thresholds (TEL, TEC, LEL, and SQS) was compared to the Level 1 and 2 biological effects levels, and the higher of the two thresholds (PEL, PEC, SEL, and CSL) was compared to the Level 3 biological effects levels, consistent with the narrative intent of these SQVs." The PEL and PEC SQGs should be compared to all three biological effect levels to be consistent with the data used in their derivation and their narrative intent.*" Consistency with the narrative intent for paired guidelines would preclude calculating reliability based on a single threshold. The TEL-type thresholds should be evaluated for their reliability in predicting the lack of toxicity and the PEL-type thresholds for their reliability in predicting toxicity.

References

Field LJ, MacDonald DD, Norton SB, Ingersoll CG, Severn CG, Smorong D, Lindskoog R. 2002. Predicting amphipod toxicity from sediment chemistry using logistic regression models. Environ Toxicol Chem 21(9): 1993-2005.

Field LJ, MacDonald DD, Norton SB, Severn CG, Ingersoll CG. 1999. Evaluating sediment chemistry and toxicity data using logistic regression modeling. Environ Toxicol Chem 18:1311-1322.

Ingersoll, C. G., D. D. MacDonald, et al. (2001). "Predictions of sediment toxicity using consensus-based freshwater sediment quality guidelines." Archives of Environmental Contamination and Toxicology **41**(1): 8-21.

MacDonald, D. D., C. G. Ingersoll, et al. (2000). "Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems." Archives of Environmental Contamination and Toxicology **39**(1): 20-31.

U.S. EPA (2005). Predicting toxicity to amphipods from sediment chemistry. National Center for Environmental Assessment, Washington, DC; EPA/600/R-04/030.

NOAA appreciates the opportunity to provide these comments. Please let me know if you have any questions.

Sincerely,

Robert Neely
NOAA Coastal Resource Coordinator

cc: Alyce Fritz, NOAA / NOS / ARD (by email)
Mary Baker, NOAA / NOS / ARD (by email)
Rob Gouguet, NOAA / NOS / ARD (by email)
Nancy Munn, NOAA / NMFS / HCD (by email)
Nick Iadanza, NOAA / NOS / ARD (by email)
Katherine Pease, NOAA/GCNR (by email)
Chip Humphrey, USEPA (by email)
Eric Blischke, USEPA (by email)
Joe Goulet, USEPA (by email)
Chris Thompson, Environmental International (by email)
Rose Longoria, Confederated Tribes and Bands of the Yakama Nation (by email)
Jennifer Peterson, Oregon Department of Environmental Quality (by email)
Jeremy Buck, USFWS (by email)